DOCUMENT RESUME

ED 224 837 TM 830 031

AUTHOR Choppin, Bruce

TITLE A Two-Parameter Latent Trait Model. Methodology Project.

INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.

SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 82
GRANT NIE-G-80-0112
NOTE 37p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; Algorithms; *Guessing (Tests); Item Analysis; *Latent Trait Theory; *Mathematical Models; *Multiple Choice Tests; Probability; Test Construction; Testing Problems; Test Items
IDENTIFIERS *Item Parameters; Two Parameter Model

ABSTRACT

On well-constructed multiple-choice tests, the most serious threat to measurement is not variation in item discrimination, but the guessing behavior that may be adopted by some students. Ways of ameliorating the effects of guessing are discussed, especially for problems in latent trait models. A new item response model, including an item parameter to describe guessing is presented. Rather than estimating the asymptotic probability for success for a person of infinitely low ability, this parameter is shown to indicate the location on the ability scale below which guessing may be anticipated to be dominant for any item. Multiple choice item data are examined to establish typical shapes for item characteristic curves and to identify possible reasons for their unanticipated variation in the lower ability scale. The model is viewed as the sum of two logistic functions: the classical Rasch model of correct response probability, and an added function where the maximum success probability is constrained by item format (number of alternative choices) and the probability that an individual will choose to guess at random in inverse relation to the person's ability. An estimation algorithm for improved item calibration and person measurement is presented. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED224837

Deliverable - November 1982

METHODOLOGY PROJECT

A TWO-PARAMETER LATENT TRAIT MODEL

Bruce Choppin
Study Director

Grant Number
NIE-G-80-0112, P3

TM 830 031

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

"It is because true measurement is essential to the discovery of laws that it is of such vital importance to science."
(Campbell, 1952, p. 134)

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

I. INTRODUCTION AND OVERVIEW

The basic premise of this paper is that on well-constructed multiple-choice tests the most serious threat to measurement is not variation in item discrimination but the guessing behavior that may be adopted by some students. Since multiple-choice tests came into widespread use in the 1920s there has been a steady stream of research studies aimed at finding ways of ameliorating the effects of guessing. Many of the early papers were based on a simple model that said that if a candidate knew the correct answer to a question he would choose it; if not he would omit the item or choose at random among all the alternatives presented. This permits an estimate of the number of items on which guesses have been made:

$$G = \frac{mW}{(m-1)}$$

where W is the number of incorrect alternatives selected and m is the number of alternative choices per questions. Assuming that $\frac{1}{m}$ of the guessed responses are correct, this suggests that the subtraction of $\frac{W}{(m-1)}$ from the raw score, R, would remove the inflation caused by lucky guesses. This, the so called "standard correction" for guessing, has come into widespread use. (It should be noted that the same principle can be applied to items rather than to persons in order to estimate the number of candidates who can genuinely solve an item.)

This standard correction has been attacked ever since it was first introduced. Its assumptions are too simple to be credible. In general a student who does not know the right answer may still know enough to be able to eliminate one or more of the distractors so that, when he comes to guess, his probability of success would be greater than $\frac{1}{m}$ (Little & Creaser, 1966). This would suggest a higher proportion of correct guesses so that, in general, the standard guessing correction would be too small. However, empirical studies (e.g., Ruch & Stoddard, 1925; Brownless & Keats, 1958) suggest that the standard correction is too large. Choppin (1974) demonstrated that a significantly smaller correction could lead to increases in both reliability and validity.

Guessing poses special problems for latent trait measurement models. The Rasch model was developed for (and initially applied to) the analysis of multiple-choice tests, yet it makes no provision for guessing behavior. The model implies that the probability of an individual responding correctly to an item tends towards a limiting value of zero as the difficulty of the item increases. Most users of the Rasch model have recognized that this is an unrealistic assumption, and some (e.g., Traub & Wolfe, 1981) have concluded that in consequence the Rasch model is not appropriate for the analysis of multiple-choice tests.

On the other hand, Lord's work on a three-parameter logistic model, which led to the development of his "Item Response Theory", recognized the occurrence of guessing from the beginning and sought to model it by using an item characteristic curve with a built in, non-zero, lower asympotote. This model includes an item parameter, c_i (that used to be known as the guessing parameter) which represents the probability of a successful

response on item i for a person of infinitely low ability. Unfortunately though this goes some way toward reproducing the form of item characteristic curves constructed from real data, it does not help in the measurement of individuals. The c_i parameter affects the probability of a correct response by every person to every item in a way that defies logical interpretation, and although the parameter contains information about the rate of successful response for people of very low ability, it does not lead to more accurate estimations of ability for persons at any level.

This paper will introduce a new item response model based on the Rasch model, but including an item parameter to describe guessing. Rather than estimating the asymptotic probability for success for a person of infinitely low ability, this parameter will indicate the location on the ability scale below which guessing behavior may be anticipated to be dominate for any item. This suggests a method for its elimination from procedures of item calibration and person measurement.

The paper consists of three parts:

(i) An examination of multiple choice item data from samples of students in several countries, in order to establish typical shapes for item characteristic curves, and to identify some posible reasons for their unanticipated variation in the lower portion of the ability scale.

(ii) An alternative model with two item parameters will be proposed, its characteristics illustrated with simulated data, and its analytic possibilities explored.

(iii) The practical application of this approach will be discussed, and an algorithm for improved item calibration and person measurement will be presented.

II. ITEM CHARACTERISTIC CURVES

Item characteristic curves relate the ability of a person attempting a test item (scaled along the horizontal axis) to the probability of that person responding correctly to the item. The use of these curves has increased primarily because of the heightened interest in latent trait theories of measurement. Most of the curves that have been published in the recent professional literature derive from some theoretical latent-trait model rather than from simple tabulations of raw data.

However, if this practice becomes standard, then there is a real danger that the truly aberrant characteristics of some items will be overlooked. ICC's drawn after fitting an item as well as possible to the preferred model appear much more reasonable than those drawn directly from raw data. In general, such curves may be dubbed "unrealistic characteristic curves" (UCC's). They bear the same relationship to true characteristic curves as does the mathematician's "smooth light frictionless pulley" to the sort you can buy in a hardware store.

While a number of different latent-trait models based on the normal or logistic ogive have been proposed, only two are in widespread use by test developers and measurement specialists. These are the three-parameter logistic function developed by Lord (1980) in his work on item response theory, and the one-parameter logistic function developed by Rasch (1960, 1980). Although different models imply different curves, certain UCC features are common to most or all models. The curves are monotonic increasing functions of ability, i.e., the probability of a successful response to an item increases as the ability of the person attempting it

increases. Secondly the curve has ogival form with the probability of success asymptotically approaching 1.0 as ability increases and asymptotically approaching a lower bound (which may or may not be zero) as ability decreases.

The Rasch function is mathematically simple and its use as a base for building a test theory has some very attractive features. According to the model, item characteristic curves differ only in a horizontal displacement corresponding to the difficulty level of the item, and curves for a group of items take on a quasi-parallel form, as shown in Figure 1a. By

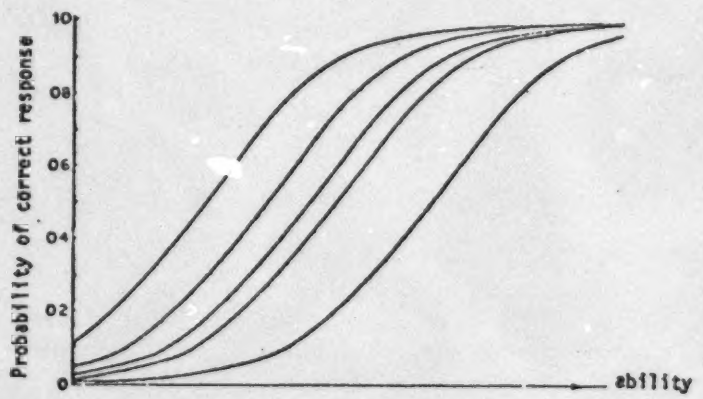


Figure 1(a)

Item Characteristic Curves for the Rasch Model

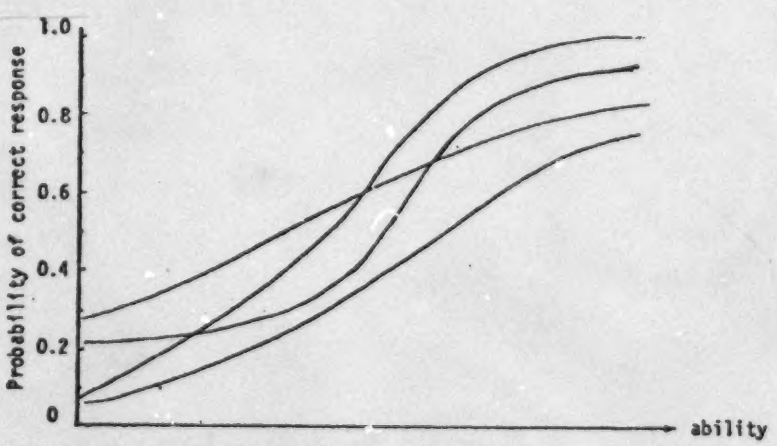


Figure 1(b)

Item Characteristic Curves for the 3-parameter Model

contrast, the three-parameter model allows for ogives which intersect each other, and which have different lower asymptotes. It has been argued persuasively that these characteristics, representing as they do varying discrimination powers of items and the possibility of responding correctly through random guessing, more truly represents the behavior of items in the real world. Advocates of the Rasch approach contend that the occurrence of random guessing, and more especially the crossing of item characteristic curves, are not desirable properties of a measurement system, and indeed are characteristics to avoid if at all possible. The purpose of this paper is to argue that neither form of curve may provide an adequate description of a multiple-choice item's behavior over the full range of ability, and to suggest ways in which such imperfect items may still be exploited to yield better measurement.

Plotting Realistic Characteristic Curves

As noted above, most of the previous literature about item characteristic curves has been based on the theory of some mathematical model. What has typically been done to use a set of item response data to estimate the parameters of a chosen model. Then a smooth curve is plotted using these parameters and the function specified by the model. The curve thus produced is not so much a summary of the raw data as an illustration of the best fitting function to a particular set of data for a chosen model (i.e., family of functions). Thus even when working from the same set of

item response data, Rasch modelers and three-parameter modelers will derive different UCC's for the same item.

What has been little attempted is the study of item characteristic curves directly, and without invoking any particular model. This may appear surprising but there are at least three good reasons for it. One is that, in the absence of a specific model, the ability scale which forms the horizontal axis of the graph is not clearly defined. Raw scores on a test or percentile scores have been used, but both are subject to serious drawbacks. Secondly, even if measures of ability are available, it is not possible to observe the probabilities for responding correctly in a direct fashion. Rather one must group together a substantial number of people at a particular level of ability, and use the relative frequency of success on the item among this group as an estimate of the probability of success for a single person. To estimate probabilities in this fashion, large numbers of subjects are needed,and this give rise to the third difficulty. Even with samples of several thousand students it is difficult to obtain sufficient data near the low end of the ability scale to plot characteristic curves with sufficient accuracy, although this region is the most contentious between the advocates of alternative models. It would be possible to design a study in which large samples of students were exposed to test materials known to be "too difficult" for them, but there is an understandable reluctance on the part of educators to subject their students to such discouraging experiences in the cause of what many will see as an intangible and esoteric research endeavor. Hence, as suggested above, direct plotting of item characteristic curves is rare.

Source of Data for the Present Study

The analyses reported below are a by-product of a research study originally conducted for quite different reasons. In 1971 the IEA carried out an international survey of science education in 19 different countries. In the course of this survey, representative samples of several thousand students in each country were given carefully constructed multiple choice achievement tests. Because of the need to obtain adequate measures over a wide range of achievement levels, over varying science curricula, and indeed over wide ranges in the availability of science teaching facilities within the participating countries, the test construction process was protracted and difficult. In the end the test instruments that were developed were quite lengthy and represented a compromise between the curricula emphases of the different countries. All the items were pretested extensively so that, in general, the psychometric quality of the instruments produced was felt to be very high (Comber & Keeves, 1973).

The results presented below relate to the tests given to students in the final year of secondary education (grade 12 in the United States, but the exact definition of the population varied according to national circumstances). All students within the chosen sample took a multiple choice test of 60 items. I have used the performance on the first 50 of these items to provide a criterion measure of the students' ability and have then used this to explore the characteristics of the remaining 10 items, which included several that were comparatively difficult. I have used the data from those countries where the sample of students tested was large, and where there was strong internal or external evidence that the fieldwork had been well conducted.

It is clearly important that the criterion measure of ability should not be contaminated by the items whose characteristics are the subject of study. Under the Rasch model, the raw score on the 50-item criterion test is a sufficient statistic for ability, and preliminary analyses show that the Rasch model provided a reasonably good representation of the data for the 50-item test. In figures 2 through 5, therefore, the horizontal axis is calibrated in raw score on the 50-item test scaled in accordance with the Rasch model. It is apparent that the use of other models would slightly vary the scaling on this axis and might prevent calibration in raw scores, but it would not substantially change the shapes of the curves or the results to be reported.

Since all 50 items were of the five-way multiple choice variety, and students may well have employed random guessing for at least some of their responses, raw scores of ten or below cannot be taken as reliable indicators of the student's ability. However, by concentrating the investigation on "difficult" items, it was possible to explore the lower reaches of the item characteristic curves, while still in ability regions where the criterion measure was adequate.

Results

Figures 2 through 5 present item characteristic curves for four selected items in five different countries. These items were chosen to illustrate the range of behavior found among the ten items investigated in detail. The countries included two for which the items had been translated into something other than their original English. The curves are plotted over a range of criterion scores for which the sample sizes were adequate (even though criterion scores below 10 have little meaning). There were

Figure 2

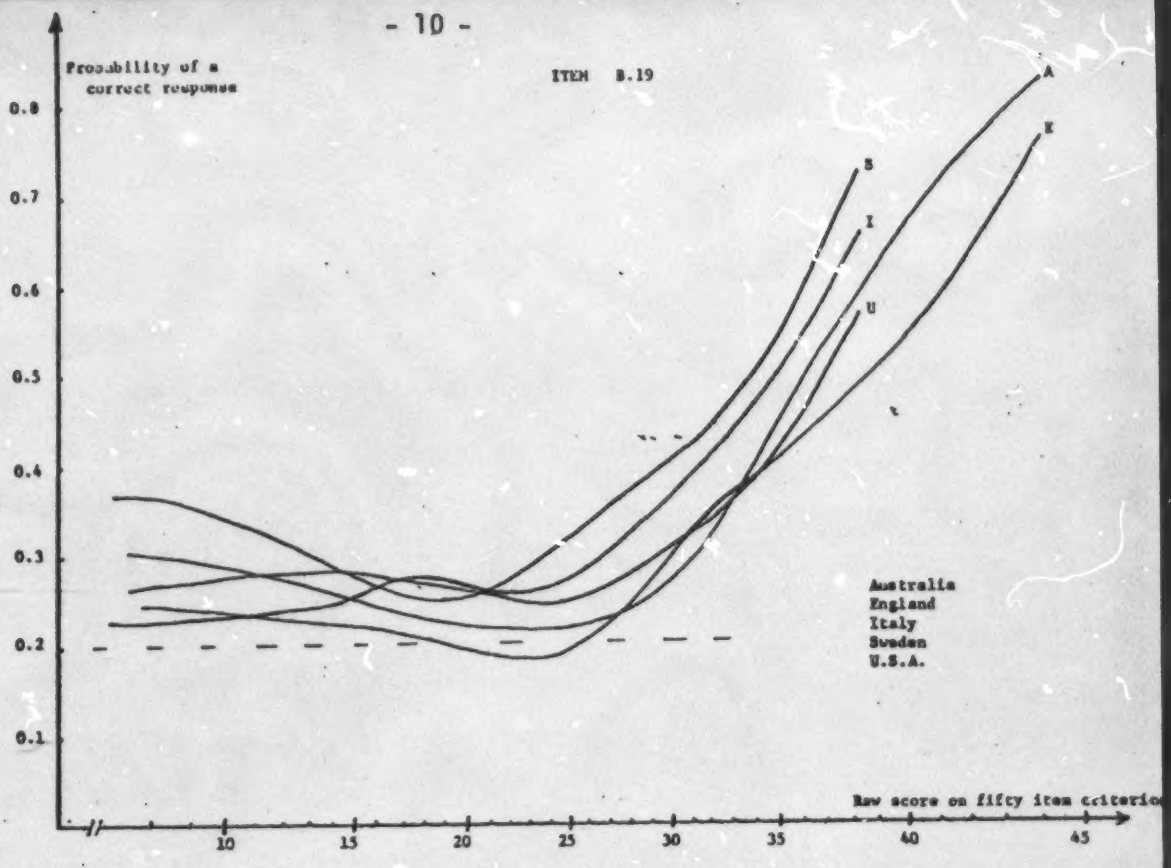


Figure 3

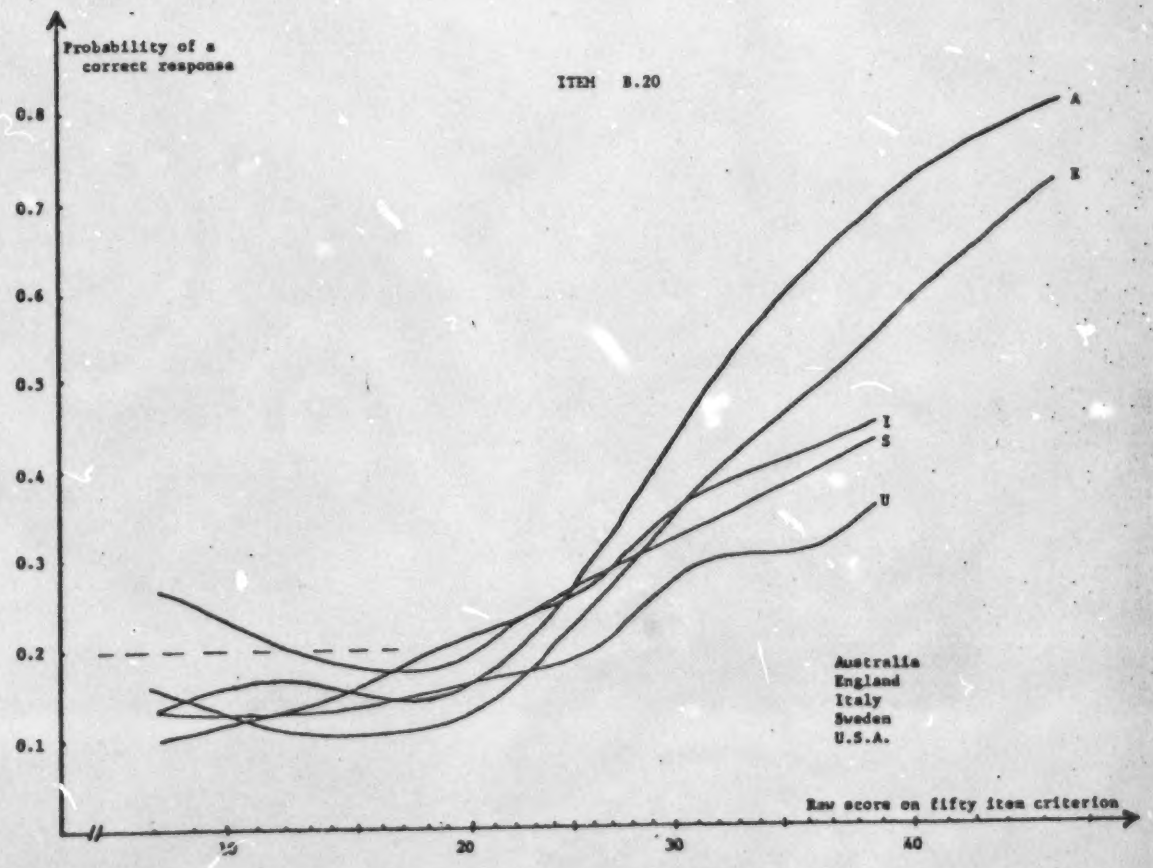


Figure 4

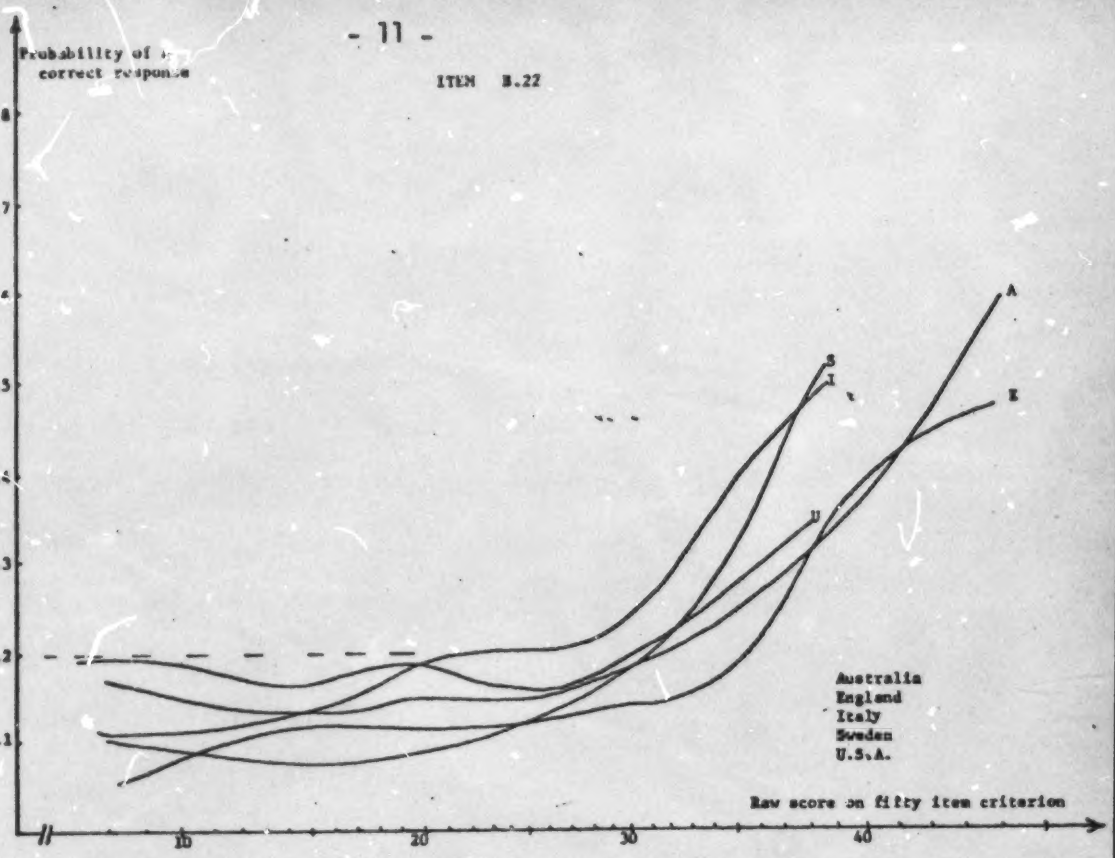
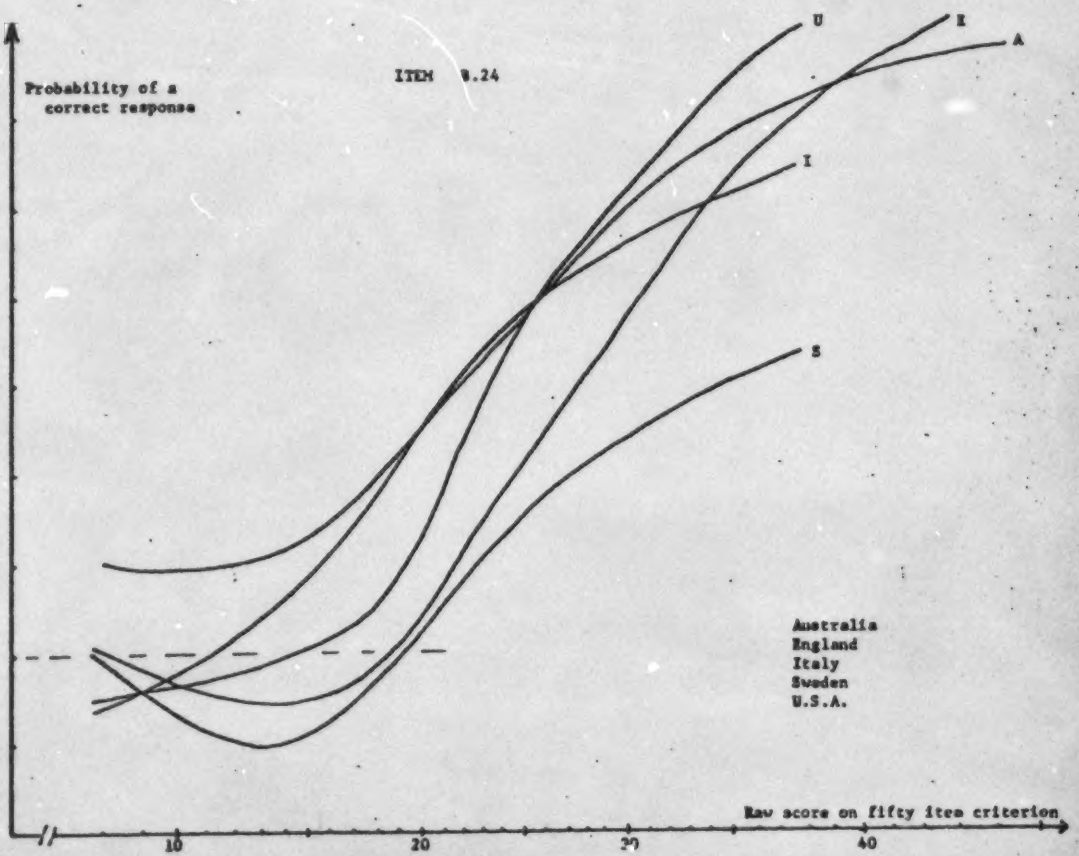


Figure 5



too few students in the USA, Sweden and Italy samples to plot curves beyond criterion scores of 38. The cut-off score for Australia and England was 44. The curves have been smoothed by local fitting to cubic polynomials. In each figure a dotted line corresponding to a probability of 0.2 has been drawn to represent the success rate of a simple random guessing strategy.

The following two points about the curves may be made:

a) The item curves reveal considerable similarity between countries, but there are some notable discrepancies. In part these may be attributed to the difficulty of obtaining exact translations (even within the three English speaking countries, some words carry slightly different meaning), but the main source of the discrepancies is almost certainly the differences between the curricula. This was to be expected, and does not in itself invalidate the international study. While individual items might betray "bias" between countries, the compromise represented by the total test was designed to be fair to all. Nevertheless, the dangers of interpreting results of single items in cross-cultural studies are well illustrated here.

b) There is strong evidence that the characteristic curves of some items do not possess the monotonic increasing property of an ogive required for scaling and measurement. If there is not a clear one-to-one correspondence established in the mapping of

observations onto the set of numbers that are to be used as measures, then the procedure is not one of measurement (Kaplan, 1964). Here we have evidence that for some test items the probability of success may actually increase as we move downwards through some parts of the ability scale. This appears to be connected to (but is not entirely explained by) the tendency to guess at random (a behavior that may be attractive to persons of very low ability).

Distractor Analysis

The J-shaped nature of some item characteristic curves has been noted before, and it has been suggested that this results from some distractors being "too attractive" to certain students. As Bock has argued, these distractors actually distract. To explore the matter further, characteristic curves were plotted for the four items already presented and, Figures 6 to 9 give the results. Now the vertical axis, instead of representing the probability of a correct response, represents the probability of choosing a particular response alternative. The data in Figures 6 to 17 are from the USA, England and Australia samples. Distractor behavior for the samples in Sweden and Italy has not yet been examined.

B.19 A certain force was needed to keep a trolley moving along a horizontal surface at a uniform velocity because the trolley had

a. inertia.
b. weight.
c. friction forces equal to this force.
d. friction forces just less than this force.
e. mass.

In Figure 6-8 it is apparent that item B.19 is highly discriminating for people whose ability puts them in the 30+ score range, but for people with raw scores below about 25 the characteristic for response c (the correct response) has a negative slope. In Australia, it would appear that students of low ability (raw scores of 15 and below) can mostly eliminate alternatives a, b and e, and distribute their guessed responses fairly evenly between c and d. Reference to the question itself suggests that this is appropriate "test-wise" behavior. The pattern is less well established in USA and England. However, in each country for people with scores in the range 15-30, incorrect alternative d was the preferred response. The J-shaped item characteristic curves presented in Figure 2 result from superimposing an "attractive distractor" on an otherwise highly discriminating item.

Figure 6.

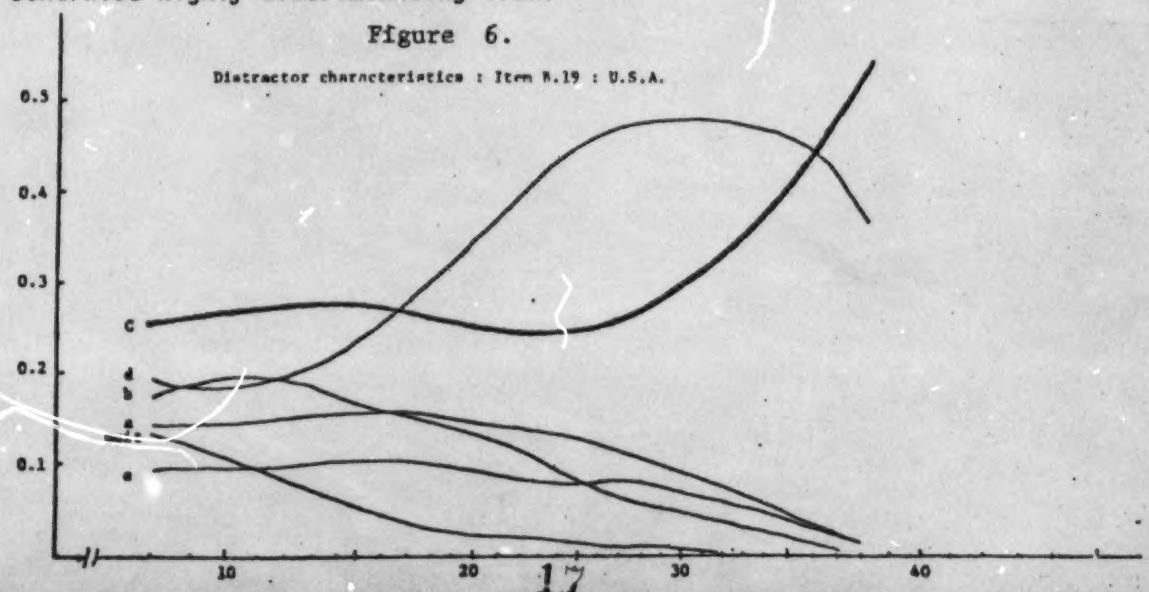


Figure 7.

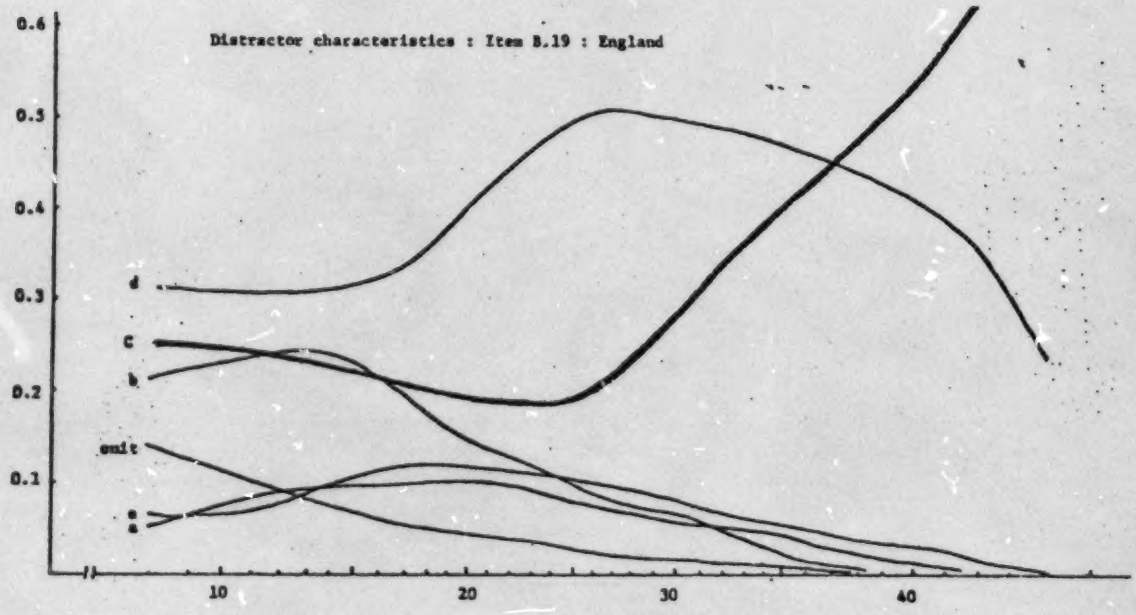
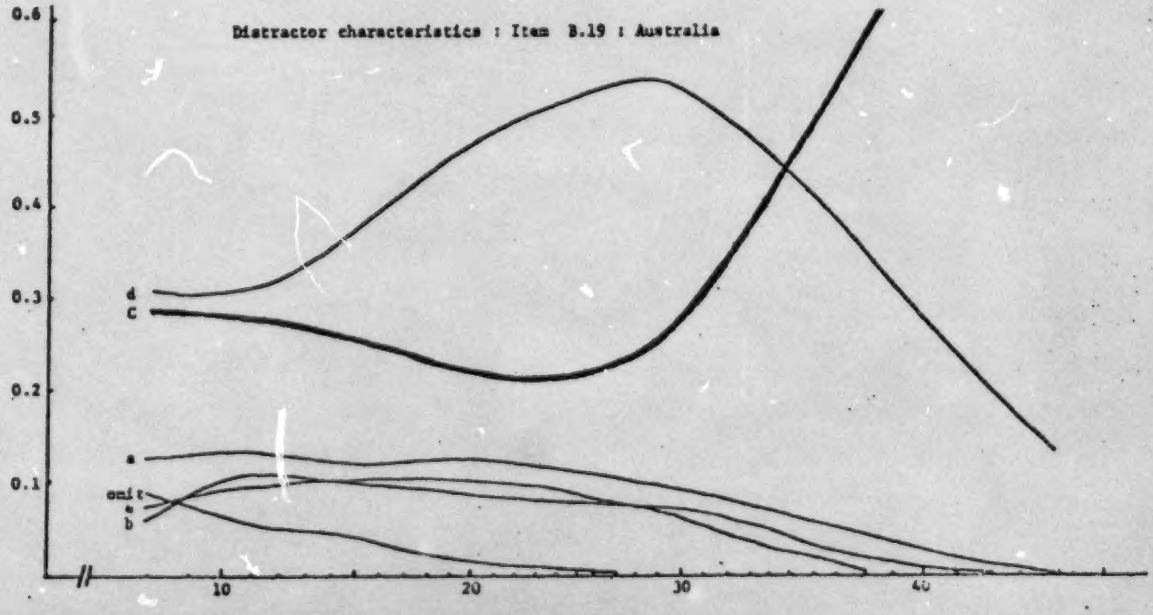


Figure 8.



B.20 A stone is thrown upward at an angle of 45°. At the highest point reached by the stone its

a. acceleration is zero.
b. acceleration is at a minimum, but not zero.
c. total energy is at a maximum.
d. potential energy is at a minimum.
e. kinetic energy is at a minimum.

Item B.20, whose distractor characteristics are shown in Figure 9-11. follows the same trend, although in this case the attractive distractor a is less dominant. Distractors b, c and d all work as expected, in Australia, but b is unexpectedly popular in the USA. The attractions of response a for people with scores in the middle region are sufficient to produce slight kinks in the characteristic for the correct response e . Science educators might well argue that choosing a betrays a complete misunderstanding of the concepts tested by the item, and that B.20 should be a good discriminator for all students. The curves suggest, on the contrary, that this item contributes no useful measurement information for students with raw scores below about 22 (26% of the sample) in Australia, the country in which it appears to work best.

Figure 9

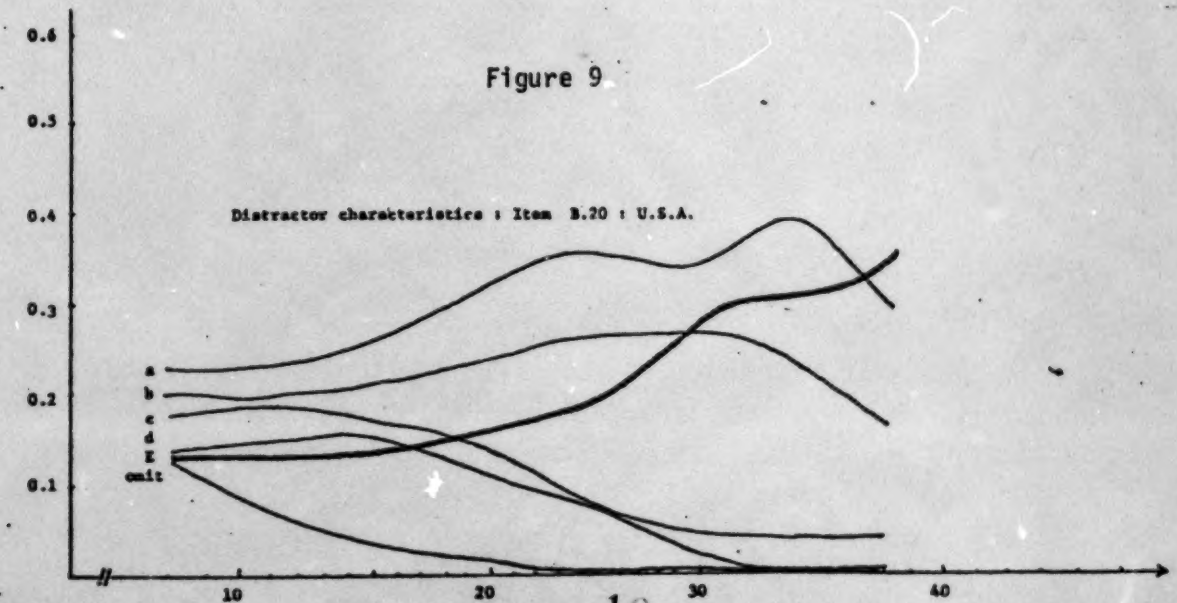


Figure 10

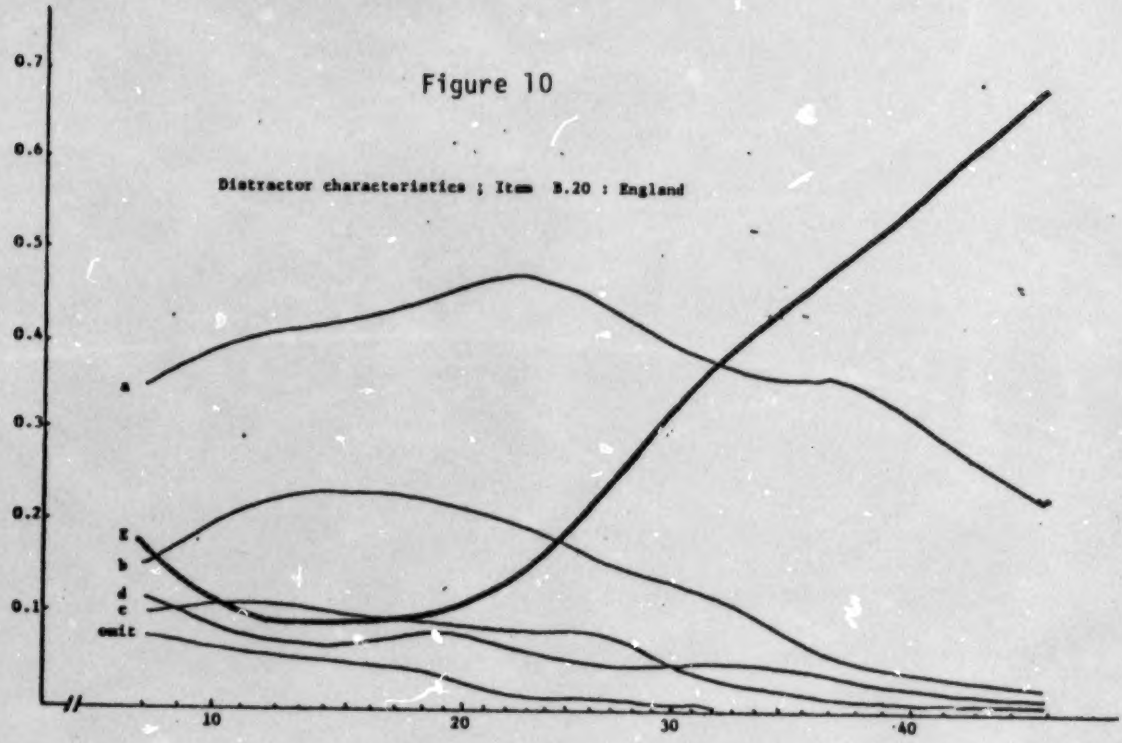
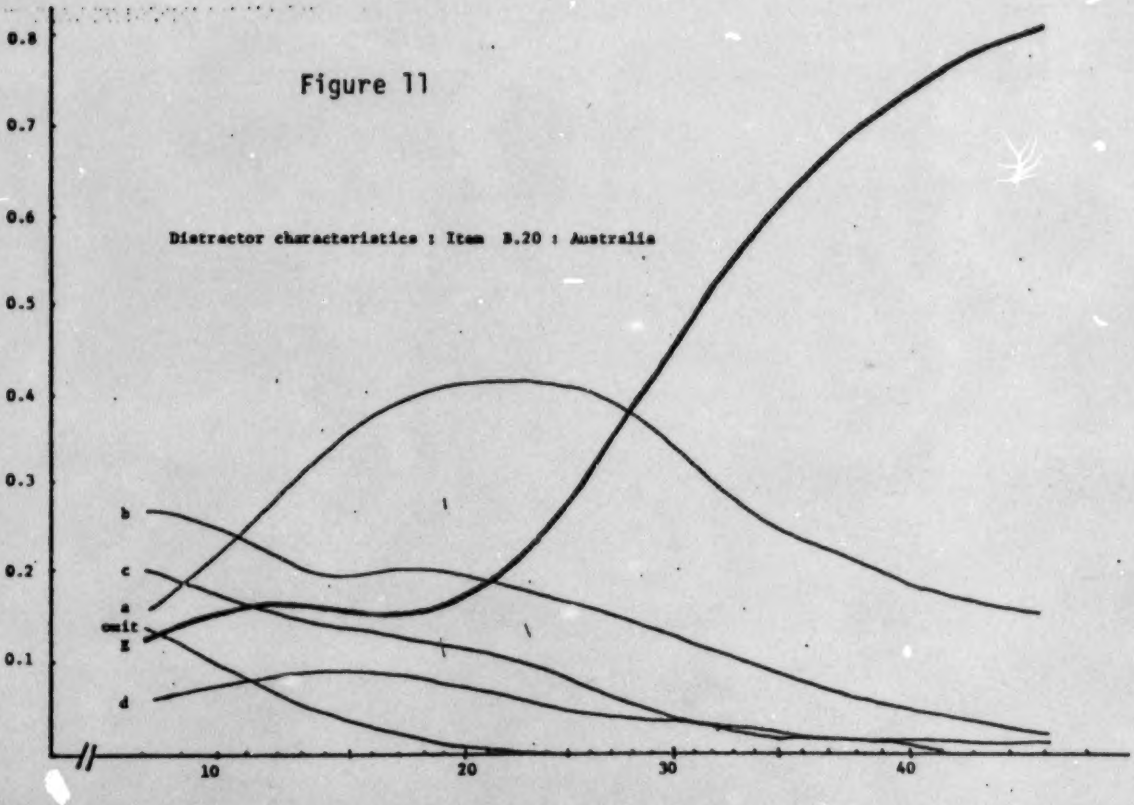


Figure 11



B.22 A one-ton truck coasts from rest down an incline of a vertical height of 30 meters and is breaked to a stop at the bottom. Air friction is negligible. In order to estimate the quantity of heat produced what additional information is required?

a. The length of the incline.
b. The length and slope (gradient) of the incline.
c. The rise in temperature of the brake surfaces.
d. The average speed of the truck.
e. None of the information in statements a to d is required.

Item B.22, presented in Figure 12-14, is unusual in that it contains two or three attractive distractors. The students of low ability overwhelmingly preferred alternative d (the "common sense" approach) while those in the 22-38 score bracket preferred c (the "scientific" answer). Response b is particularly popular in the United States. The correct response e is of the "none of the above" variety, and test-wise students usually avoid such responses. It may be noted that in the United States, where test-wiseness is probably greatest, this item barely discriminated at any point of the scale. In England it discriminated well for students who scored above 30, but of those with scores below 30 very few of the students got it right.

Figure 12.

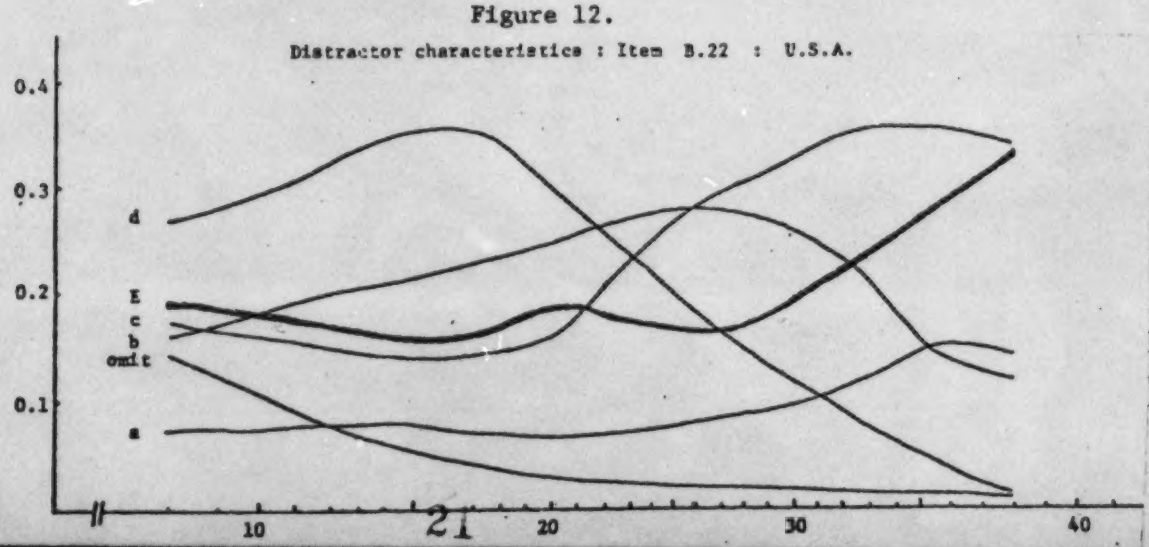


Figure 13

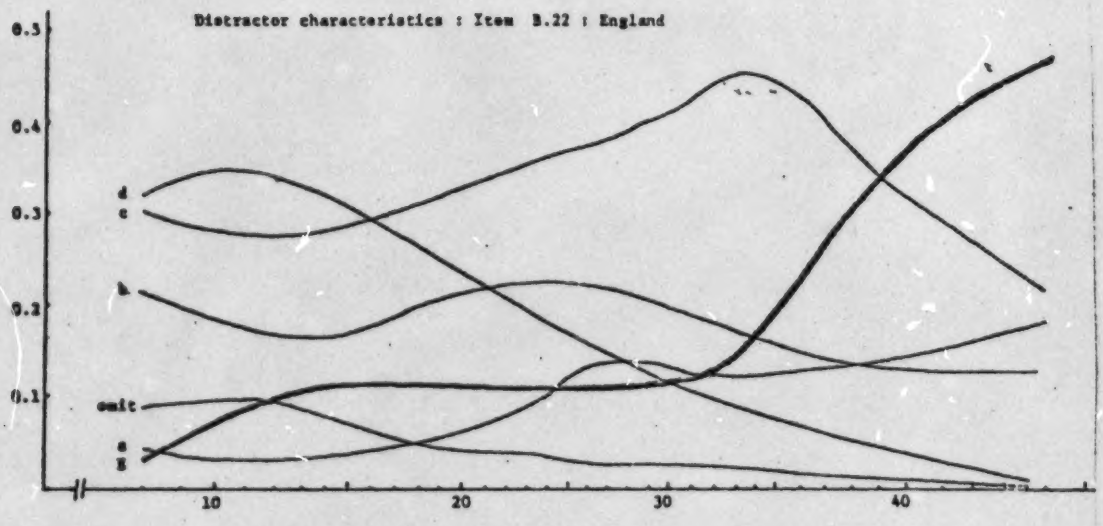
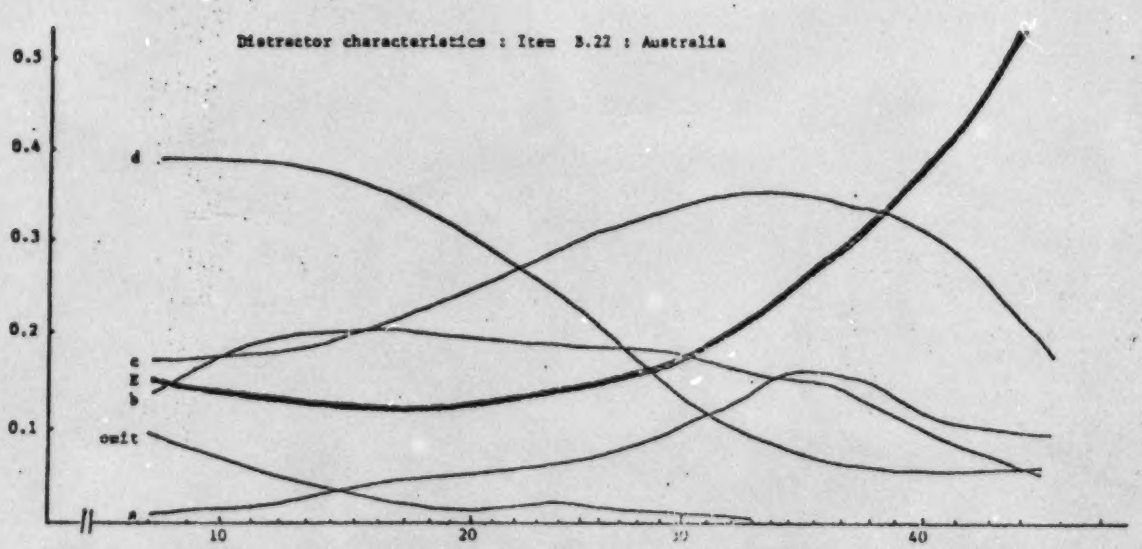


Figure 14



B.24 By which of the following methods can geological time be measured most accurately?

a. Size of fossils.
b. Thicknesses of sedimentary layers.
c. Radioactivity of uranium.
d. Rate of salt accumulation in the ocean.
e. Temperatures in the mantle.

Finally, Figure 15-17 displays distractor characteristic curves for item B.24. As can be seen from Figure 5, this item discriminated well for students of higher abilities, but displayed some peculiarities at the lower end of the scale. In Australia, the curve levels out at a much higher level of success than in the other countries. From Figure 9 it is clear that distractor b is the cause of the abnormal behavior. In Australia, it was the preferred response for students who scored 20 or below, but it is comparatively less attractive than in the other countries examined (no doubt because of the geological experiences of Australian students). In the USA, a student with a score of 15 was three times more likely to select

Figure 15.

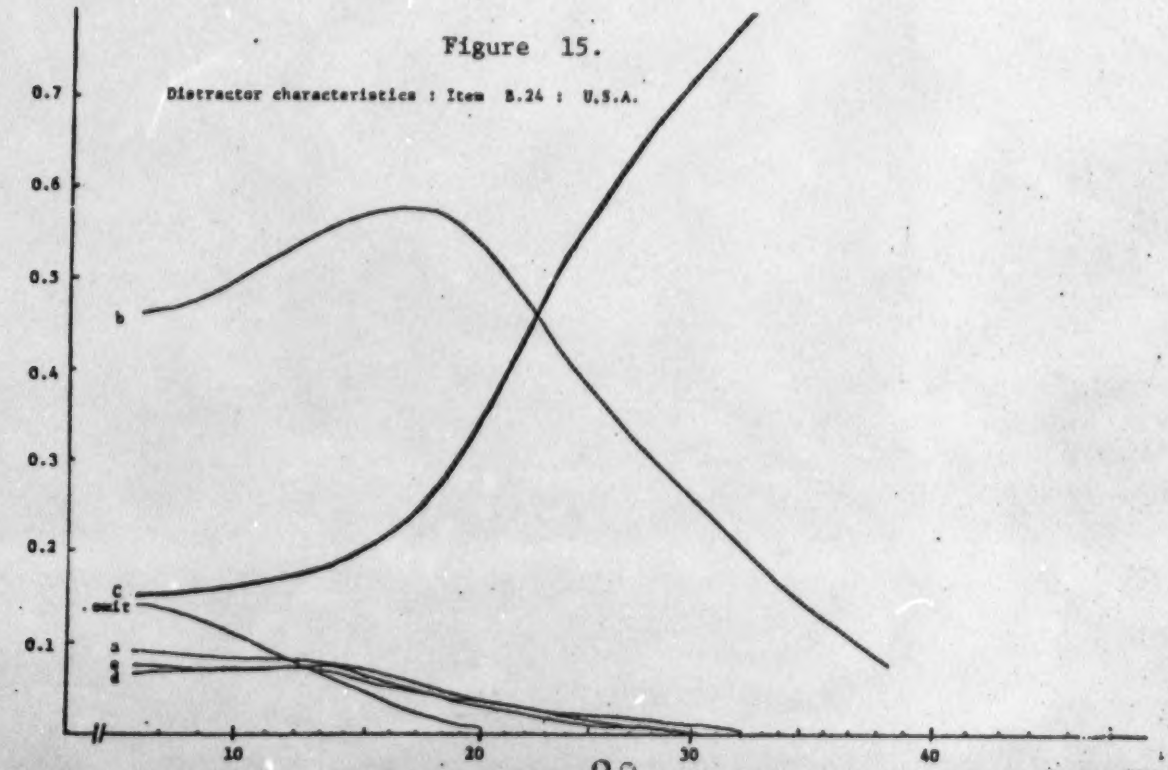


Figure 16

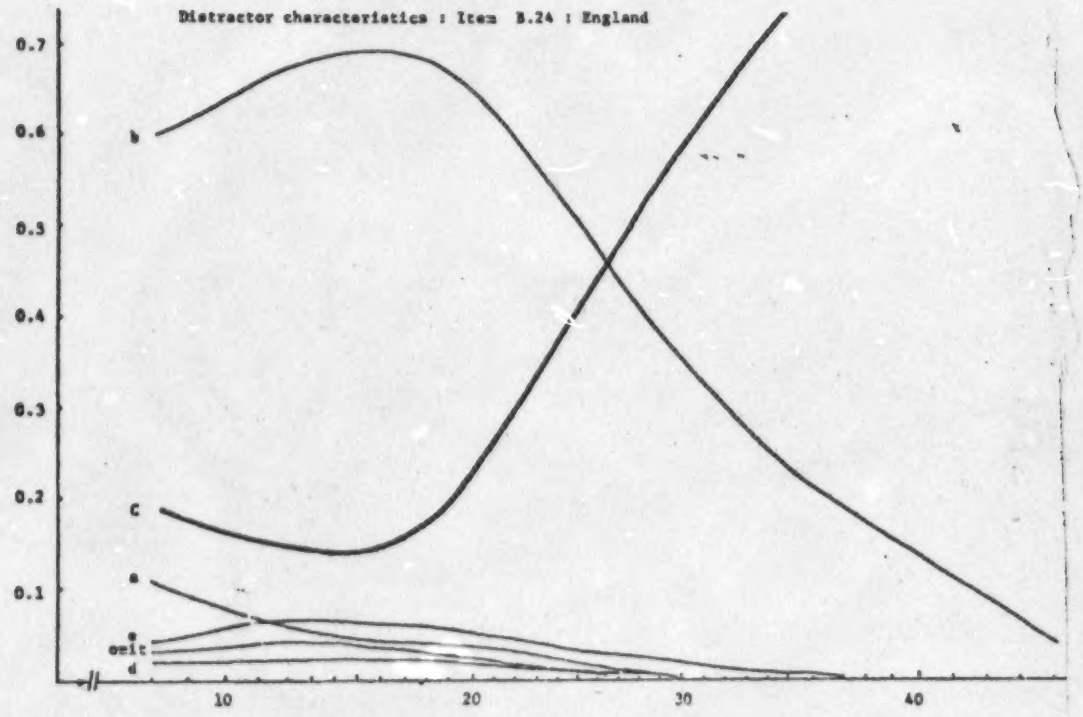
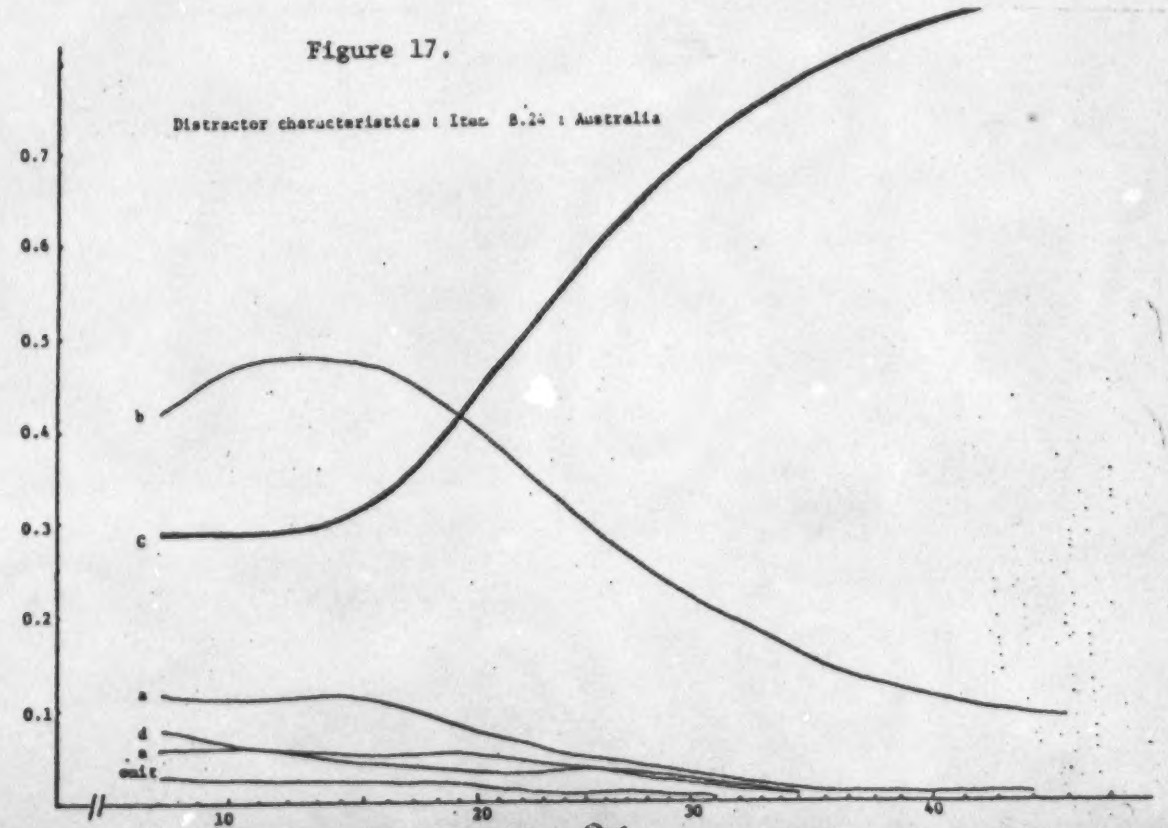


Figure 17.



response b than to choose the correct answer c , while in England the comparable ratio was more than 4 ($p = 0.68$ against $p = 0.16$). The other distractors play little role in this item (except for option a at the very bottom of the scale). This is an item that works effectively for raw scores above 15 where the preference for response c over b clearly discriminates according to level of achievement. The item is not contributing to the effective measurement of persons with raw scores less than 15.

Where the monotonic increasing nature of the characteristic curve is not firmly established there is doubt about the validity of incorporating results on such an item into a score which is used to measure a student's ability. All ten of the items investigated in this study, and not only the four whose characteristic curves are presented here, are good and effective instruments of measurement for part of the ability range, but completely ineffective in other parts. The J-shaped non-monotonic nature of the item characteristic curve is firmly established for four of these ten items and is a tenable hypothesis for three others.

Further it is clear that neither the one-parameter Rasch model nor the three-parameter latent-trait model developed by Lord, are appropriate descriptions of those items where "attractive distractors" are present. The effect of these distractors on the item response function is to introduce an anomoly analogous to that found when the density of water is used to measure temperature. Unlike most other liquids, water reaches its greatest density above its freezing point so that if cooling continues down past 4°C it begins to expand again. It was found early in the history of

thermometry, but only after a good deal of head-scratching, that water was not a good liquid to use in thermometers. Arguments that, for example, "the whole notion of temperature is misconceived" faded away once it became apparent that thermometers filled with other liquids worked satisfactorily.

Now test constructors seem to be faced with a choice; either to construct items without harmful "attractive distractors", or to devise a way of using existing items only for those parts of the ability scale at which they function adequately. The remainder of this paper addresses the second possibility.

III. A NEW TWO-PARAMETER LATENT TRAIT MODEL

In developing a new latent trait model to encompass the guessing behavior we have found to occur on multiple-choice test items, it would seem appropriate to concentrate on changes that are chiefly significant at the lower end of the ability scale. Both the standard three-parameter model and the much simpler Rasch formulation appear quite adequate to describe the behavior of real items for those people for whom the item is comparatively easy. The requirement for a new model is that it should constrain the item characteristic curve to approach a fixed lower asymptote specifiable from the structure of the item (e.g., a multiple-choice item with four alternatives would require a lower asymptote of $p = 0.25$), but to do so

(a) in a continuous fashion

(b) to allow some flexibility in the shape of the ICC

(c) to minimize disturbance to the model of behavior for high ability persons.

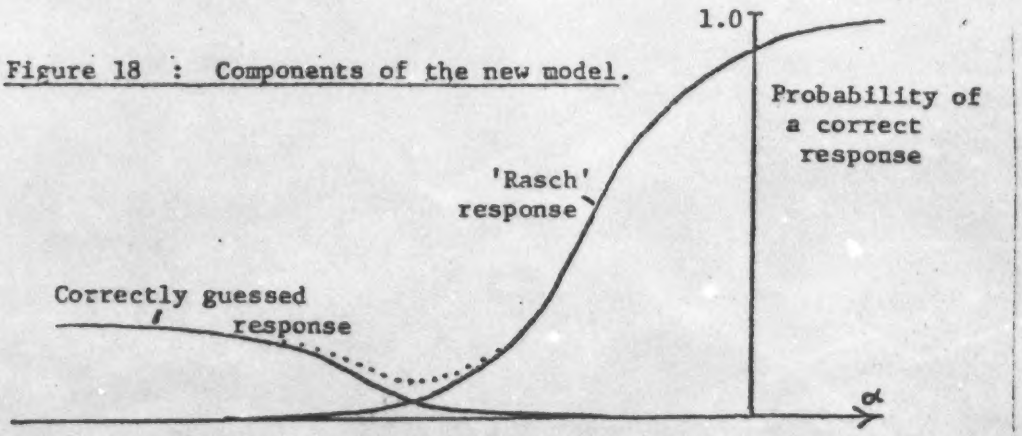
For simplicity, we shall consider modifications to the Rasch logistic model that themselves take a logistic form. It may be that other types of continuous function might yield a better fit in the case of particular individual items but the greater general utility of such a model for measuring a person's achievement would need to be demonstrated in order to justify the additional algebraic complexity. The proposed new model for multiple-choice items is:

$$\text{Prob}\left\{X_{vi} = 1\right\} = \frac{W^{\alpha_v}}{W^{\alpha_v} + W^{\delta_i}} + \frac{1}{m_i} \cdot \frac{W^{(\delta_i - \gamma_i)}}{W^{\alpha_v} + W^{(\delta_i - \gamma_i)}}$$

where m_i is the number of alternative answers provided for item i and γ_i is the guessing parameter for the same item.

It is helpful to view this new model as the sum of two simple logistic functions. The first is the classical Rasch model in which the probability of an individual responding correctly increases monotonically from zero to one in accordance with the basic Rasch formula. The second function, the one that has been added, represents the probability of obtaining the correct answer by guessing where the maximum probability of success is constrained by the format of the item (i.e., the number of alternatives between which a choice must be made) and the probability that a particular

Figure 18 : Components of the new model.



individual will choose to guess at random, which in general is inversely related to the person's ability. Thus the greater the ability of a person, the more likely he is to solve the question by normal means and the less likely is he to guess at the correct answer (see Figure 19). In general, it appears that guessing only becomes a dominant behavior for students whose ability is quite low compared with the difficulty of the item. is a measure of the distance between the points of inflection of the two curves and may be expected to vary from item to item.

Figure 19 : Characteristic Curves for an item with δ=50, m=5, and various values for γ.

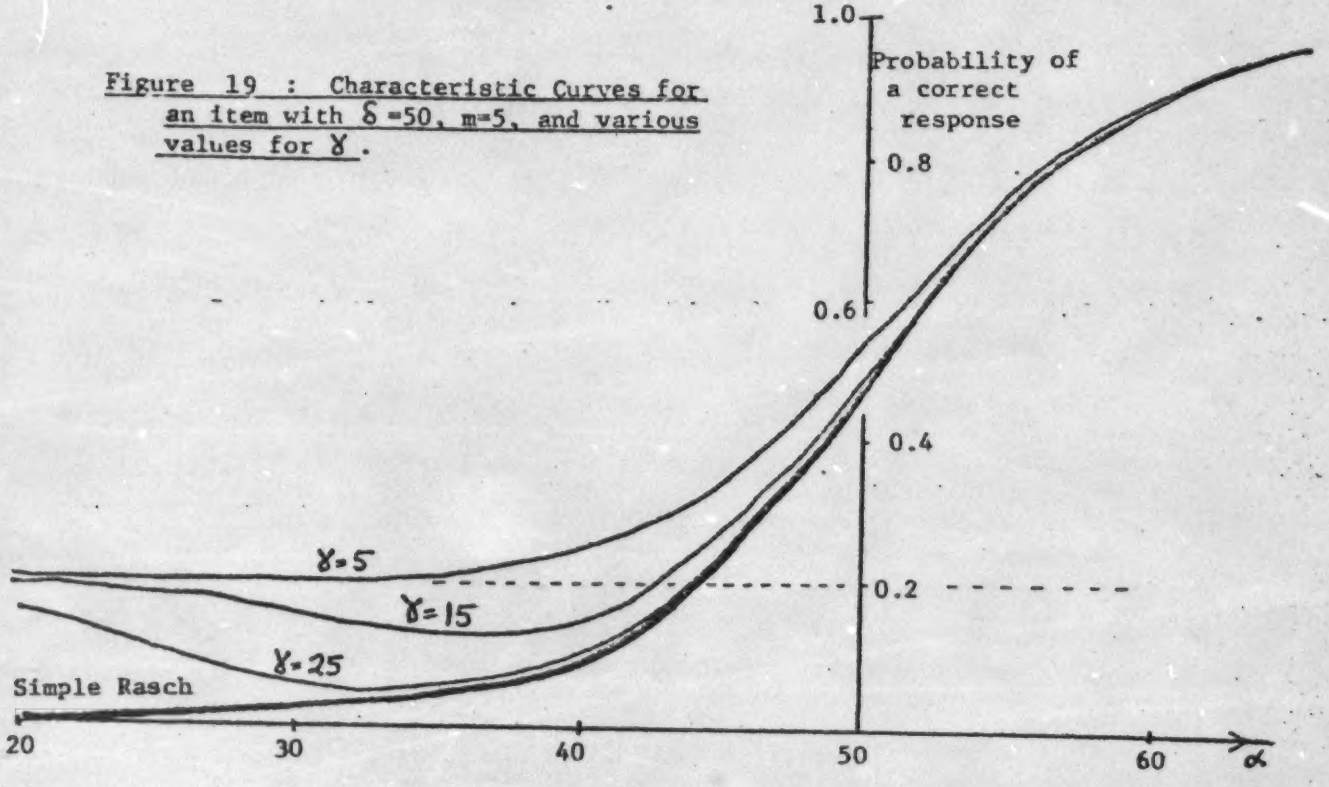


Figure 19 displays item characteristic curves under this new model for an item with m=5 and for various measures of γ . A simple Rasch UCC is plotted for comparative purposes. (The new model asymptotically approaches the Rasch model as $\gamma \to \infty$.)

In effect it can be seen that γ measures how far below the item's natural difficulty level δ one may go before guessing becomes an important factor in the student's behavior. It is clear that if students of low enough ability are included in the sample under consideration, then the new model is significantly different from the simple Rasch model for all values of γ . However, if the students being tested all have abilities of 40 wits or above, then the curves for the new model are not appreciably different from those of the Rasch model.

The discussion in the previous section suggests that the size of for a particular item is likely to be determined by the attractiveness of the different distractors to an item. Note that in the new model no attempt is made to introduce parameters for individual distractors. γ is a measure of the extent to which a student of low ability is drawn logically towards the choice of one or more attractive distractors rather than to guess at random.

Comparison of Figure 19 with Figures 2-5 suggests that reasonable γ values for the items discussed earlier in the paper are in the range of 5 to 25. However, visual comparison of item characteristic curves to some standard does not provide a practical method for estimating γ for new items because the amount of data required to construct the curves is generally not available.

An algebraic procedure for estimating γ is to be preferred, but unfortunately the usual algorithms for producing Rasch parameter estimates will not work with the new model precisely because it does not belong to the Rasch family (i.e., it is not specifically objective and its parameters

cannot be algebraically separated). For example, if we consider the joint probability of responding correctly to item i and incorrectly to item j

$$\text{Prob.}\left\{X_{vi} = 1,\ X_{vj} = 0\right\}$$

$$= \left[\frac{W^{\alpha_v}}{W^{\alpha_v} + W^{\delta_i}} + \frac{1}{m_i} \cdot \frac{W^{\delta_i - \gamma_i}}{(W^{\alpha_v} + W^{\delta_i - \gamma_i})}\right] \cdot \left[\frac{W^{\delta_j}}{W^{\alpha_v} + W^{\delta_j}} - \frac{1}{m_j} \cdot \frac{W^{\delta_j - \gamma_j}}{(W^{\alpha_v} + W^{\delta_j - \gamma_j})}\right]$$

we see that the γ's and δ's are hopelessly confounded, and the confounding of γ with δ is sufficient to prevent the elimination of α as in the standard Rasch pair-wise procedure.

If we restrict consideration to abilities that satisfy $\alpha > \delta_i - \gamma_i$, then the model can be written as a series:

$$\text{Prob.}\left\{X_{vi} = 1\right\} = \frac{W^{\alpha_v}}{W^{\alpha_v} + W^{\delta_i}} + \frac{1}{m_i} \cdot \left[\frac{1}{W^t} - \frac{1}{W^{2t}} + \frac{1}{W^{3t}} - \frac{1}{W^{4t}} \cdots\right]$$

$$\text{where } t = \alpha_v - (\delta_i - \gamma_i)$$

and if α is large enough, the equation approximates the simple Rasch model.

In practice, this may be all that is required. Though it is of theoretical interest to know about the size of γ for an item and to measure the attractiveness of its distractors, it is not necessary in order to measure person ability. What are required are adequate calibrations of the δ's (based on people whose ability is such that the guessing term may be safely disregarded) and then a set of responses by a particular person to those items judged "not too difficult" for that person in order that α may be estimated without contamination by γ .

Responses to items made by people for whom the item is difficult, and who therefore may well be guessing, are not used in the calibration of items. Responses made to difficult items, and that may well be the result of guessing, are not used in the estimation of a person's ability. The next section describes a procedure for carrying out these processes.

IV. A NEW ESTIMATION ALGORITHM

The algorithm set out below is less complex than it may appear at first sight. It consists of a sequence of Rasch scaling procedures applied to edited versions of the test data matrix. For simplicity, it will be presented as a series of discrete steps applied to a raw data matrix containing a complete set of responses from N people to a test of k items. In practice several steps might be collapsed into one, and the method can be applied to more complex data structures.

Step 1: Response Scoring

The raw data matrix is scored such that correct responses are coded 1, and incorrect responses zero. The rectangular N by k matrix thus produced is completely filled with 1's and 0's.

Step 2: Initial Calibration of Item Difficulties

The entire matrix is used to estimate δ values for the items using the PAIR algorithm. A matrix of b_{ij} elements is developed where b_{ij} is the number of people who respond correctly to item i and incorrectly to item j. Analysis of this matrix yields maximum likelihood estimates for all k δ's.

Step 3: Initial Estimation of Abilities

These estimates of δ are used to obtain α_r values for each possible raw score on the test from 1 to (k-1). Maximum likelihood estimates of α_r arise from iterative solutions to the equation

$$r - \sum_{i=1}^{k} \frac{W^{\alpha_r}}{W^{\alpha_r} + W^{\delta_i}} = 0$$

where r is the raw score being considered.

Step 4: Item Screening Table (Table A)

A table is constructed to show, for each raw score on the entire test, which items are probably not effective for measuring a person with that raw score (i.e., the items to whose calibration such a person cannot effectively contribute). This might be done by identifying all those items whose δ value is 5 or more wits greater than the corresponding ability estimate, since a priori the model predicts a success rate < 0.25 on such items.

Step 5: Final Calibration of Item Difficulties

The data matrix created in Step 1 is rescanned, one person at a time. For each person, the raw score on the complete test is used in conjunction with Table A to determine which items do not provide reliable information. Corresponding elements in the response vector are replaced by blanks (to indicate missing data). Then the vector is used to accumulate the matrix of b_{ij} values required for item calibration although this time b_{ij} is incremented only if both responses to items i and j are present. This procedure effectively removes the vast majority of the "guessed" responses from the calibration process. After scanning all the original data matrix, the resulting matrix of b_{ij}'s is used to develop the final item calibrations. For the easier items the results are similar to those obtained in the first pass, but for the more difficult items, the changes may be substantial and the standard errors may be expected to increase.

Step 6: Ability Calibration Table (Table B)

The item calibrations produced in Step 6 when combined with the item screening table (Table A), are input to the ability calibration equation given in Step 3. The result is a new table (Table B) which give ability estimates (and standard errors) for all possible raw scores on the appropriate reduced set of items for each possible raw score on the total test.

Step 7: Final Ability Measurements

Again the persons x items data matrix (as modified in Step 5) is scanned. For each person the raw score on the total test defines a subset of items for which the responses will be considered, and a reduced raw score is calculated on this subset. These two raw scores are used to identify, in Table B, the optimal estimate of ability for the person and its associated standard error. Of course the lower the raw score, or the more unexpected the pattern of response, the greater will be the standard error of measurement, but this appears to be an appropriate feature of a situation in which guessing occurs.

Rasch UCC's for an entire test take the form shown in Figure 21. There is substantial overlap between items, and the "width" of the test should be sufficient to cover the range of ability of the students for whom it is intended. However, the lower portions of the UCC's carry very little credibility in the case of multiple-choice items.

What the new algorithm is doing is estimating Rasch item characteristic curves only for probability values above a fixed criterion value. No assumptions are made about the behavior of the items at lower levels of ability. Then individuals are measured only on those items for whom the ICC's seem appropriate. The situation is illustrated in Figure 22.

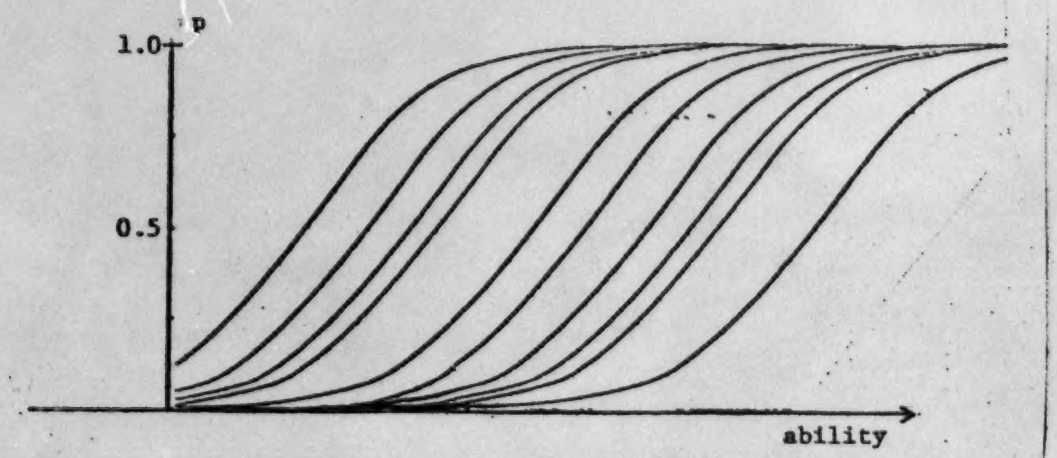


Figure 21. Characteristic curves (UCC's) for a test of ten items.

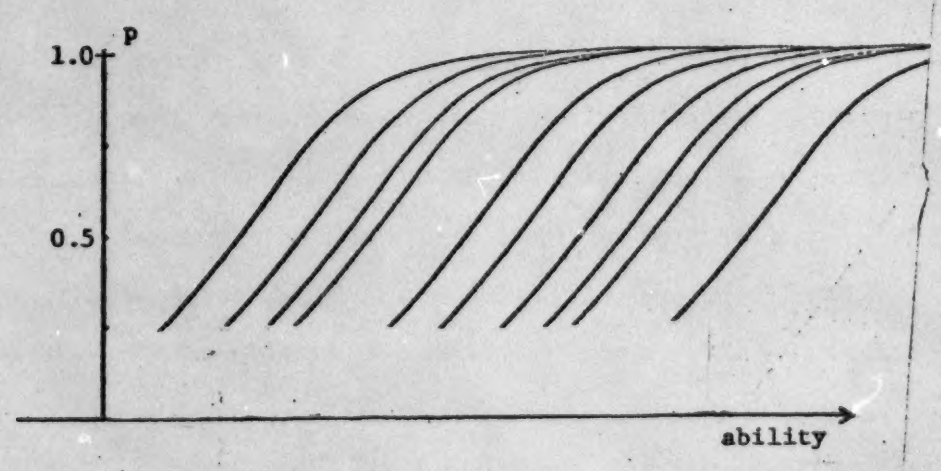


Figure 22. Modified Characteristic Curves use for ability measurement

It should be understood that this procedure works because of the capability of the PAIRWISE algorithm to provide (both least squares and maximum likelihood) item calibrations from incomplete data. This follows directly from the specific objectivity property of the Rasch model. Other estimation algorithms that handle incomplete data by dropping either items or persons completely are not useful here. The method requires that people of different abilities will be measured with different sets of items. (A strategy somewhat similar to the one above was outlined by Waller (1976), but lacking a method of estimating parameters for incomplete sets of data, he was unable to make it operational.)

For the three-parameter model which does not possess specific objectivity, the correction procedures outlined above would not work. It may be possible to develop an approximation algorithm to achieve the same result, but this has not so far been attempted, and it is not clear that anything worthwhile would be gained.

In the long term it may be better to develop test construction procedures that avoid items with "attractive distractors", or even to move away from multiple-choice to a constructed response format. Until then, a computer package to carry out the steps described above can lead to significant improvements in the measures obtained with multiple-choice tests. The increased variability in the accuracy of the estimates obtained, when compared with a standard Rasch procedure, is a nuisance, but reflects the real life situation.

The procedure outlined above has been programmed in a prototype version for use on IBM computers. Initial results support the suggestion that it can lead to substantial improvement in measurement on "difficult"

multiple-choice tests. On "easy" tests the calibration of the more difficult items is somewhat improved, but overall the changes in the ability measures assigned to persons are little changed.

The criterion level of probability values below which item responses are discounted needs to be explored empirically to establish optimum levels. Initial experiments with five-way multiple-choice items have used a criterion level of p=0.25. This value, slightly above the "chance value" of success on the item appears reasonable, but it may be possible to improve on it.

REFERENCES

Brownless, V.T. & Keats, J.A. A retest method of studying partial knowledge and outher factors influencing item response. Psychometrika, 1978, 23, 67-73.

Choppin, B.H. The Correction for Guessing on Objective Tests. Stockholm: IEA, 1974.

Choppin, B.H. Guessing the answer in objective tests. British Journal of Educational Psychology, 1975, 45, 206-213.

Choppin, B.H. Item Banking and the Monitoring of Achievement. Research in Progress Series, No. 1, Windsor: NFER, 1978.

Comber, L.C. & Keeves, J. Science Education in Nineteen Countries. Stockholm: Almqvist and Wiksell, 1973.

Kaplan, A. The Conduct of Inquiry. San Francisco: Chandler, 1964.

Little, E. & Creaser, J. Uncertain responses on multiple choice examinations. Psychological Reports, 1966, 18, 801-802.

Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum, 1980.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, 1960; reprinted Univ. of Chicago Press, 1980.

Ruch, G.M. & Stoddard, G.D. Comparative reliabilities of five types of objective examinations. Journal of Educational Psychology, 1925, 16, 89-103.

Traub, R.E. & Wolfe, R.G. Latent trait theories and the assessment of educational achievement. Review of Research in Education, 1981, 9, 377-435.

Waller, M.I. Estimating parameters in the Rasch model: Removing the effects of random guessing. Research Bulletin 76-8, Educational Testing Service, 1976.